Cleaning noisy wordnets

Benoît Sagot[1], **Darja Fišer**[2]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 175 rue du Chevaleret, 75013 Paris, France
2. Department of Translation, Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia

benoit.sagot@inria.fr, darja.fiser@ff.uni-lj.si

Abstract

Automatic approaches to creating and extending wordnets, which have become very popular in the past decade, inadvertently result in noisy synsets. This is why we propose an approach to detect synset outliers in order to eliminate the noise and improve accuracy of the developed wordnets, so that they become more useful lexico-semantic resources for natural language applications. The approach compares the words that appear in the synset and its surroundings with the contexts of the literals in question they are used in based on large monolingual corpora. By fine-tuning the outlier threshold we can influence how many outlier candidates will be eliminated. Although the proposed approach is language-independent we test it on Slovene and French that were created automatically from bilingual resources and contain plenty of disambiguation errors. Manual evaluation of the results shows that by applying a threshold similar to the estimated error rate in the respective wordnets, 67% of the proposed outlier candidates are indeed incorrect for French and a 64% for Slovene. This is a big improvement compared to the estimated overall error rates in the resources, which are 12% for French and 15% for Slovene.

Keywords: Wordnet, Wordnet Creation, Wordnet Cleaning

1. Introduction

In the past years, automatic creation of wordnets for new languages has become increasingly popular due to high cost of manual wordnet development and the success of recycling already existing language resources, such as bilingual dictionaries, Wikipedia, and parallel corpora (Agirre et al. 2002, Bond et al. 2008, Fišer and Sagot, 2008). Wordnets for more than 50 languages have been registered with the Global Wordnet Association[1], most of which have benefitted from automatic approaches.

However, the state-of-the art methods for the population of wordnets are still far from perfect, resulting in noisy synsets. This is why the goal of this paper is to propose a language-independent, corpus-based approach to detect outliers in automatically generated synsets and filter them out in order to obtain a cleaner, more useful lexico-semantic resource for human use as well as for various NLP tasks.

The work we present in this paper falls within the scope of distributional methods for detecting semantic similarity between words (Lin et al. 2003), but instead of identifying most closely related words according to the contexts they appear in, we start from a (noisy) list of synonym candidates in the form of an automatically induced wordnet.

In a way, our task is not very different from the lexical substitution framework (Mihalcea et al. 2010), with the exception that we are most interested in the bottom of the ranked list of potential synonyms. In addition, our notion of the synonym is much stricter because it is our aim to clean all the synsets in an automatically created wordnet, which is very fine-grained.

At the same time, the notion of polysemy that is of key importance for this work is translation-motivated. This means that regardless of the number of synsets a word appears in, the distinction between those senses that are lexicalized differently is the only relevant one in this work.

Our work focuses on identifying and eliminating the most obvious errors in synsets that occurred due to errors in word-alignment of parallel corpora (e.g. misaligned elements of multi-word expressions) and inappropriate word-sense disambiguation of homonymous words (e.g. assigning a valid translation of one sense of a homonymous source word to all its senses). It is precisely these errors in wordnets that have the biggest impact in NLP applications and decrease the value of the resource the most.

In this work, our approach relies on a simple hypothesis: lexemes, defined here as literal-synset pairs, tend to co-occur in corpora with other lexemes that are semantically related, as made explicit by relations between synsets in a wordnet. This is possible because when dealing with already large wordnets, such as the French WOLF or the Slovene sloWNet, this technique can provide a sufficient number of semantically related lexemes for most lexemes with a high precision (the precision of WOLF and sloWNet have been evaluated as 86% and 85% respectively).

This paper is structured as follows: in the next section we present the resources used in the experiment. In section 3 we describe the method we used to detect synset outliers and go through the experimental setup in detail. The results are evaluated and discussed in Section 4, and Section 5 concludes the paper and points towards future work.

[1] http://www.globalwordnet.org/

2. Resources used

2.1 Wordnets

The proposed approach is tested on wordnets for Slovene and French, which were both created automatically from heterogeneous resources (Fišer and Sagot 2008). They are both based on Princeton WordNet (Fellbaum 1998) and were built automatically in two stages, each using a different approach according to the resources used to extract lexico-semantic information.

The first approach focused on obtaining translations of the core vocabulary while making sure the correct wordnet senses were assigned to their Slovene/French equivalents by disambiguating it with a word-aligned parallel multilingual corpus and already existing wordnets for several languages. The complementary second approach was devised so that it tackled specialized vocabulary, which is largely monosemous and rich in multi-word expressions. Their translations were extracted from Wikipedia, Wiktionary and its related resources. When the results from both approaches were combined, they were used in a large-scale wordnet extension in which a maximum entropy classifier was trained to determine appropriate senses of translation candidates extracted from the heterogeneous resources described above (see Sagot and Fišer 2012).

The resulting wordnets have a reasonable coverage, with 76,436 and 82,721 lexemes (literal-synset pairs) respectively, filling 46,449 and 42,919 synsets respectively. The accuracy of the extended wordnets is estimated at 86% for WOLF 0.2 and 85% for sloWNet 3.0 (*literal, synset*) pairs. This suggests that there is quite a lot of noise in the extended synsets, which we will try to eliminate with the approach described in this paper. Note however that before performing the experiments described below, the WOLF, originally structured according to the version 2.0 of the Princeton WordNet (PWN), was mapped to PWN 3.0 (sloWNet already used PWN 3.0 synsets).

2.2 Corpora

In order to obtain distributional data for the words in the experiment, we used two large monolingual reference corpora. For Slovene, we used the 620 million-word corpus called FidaPLUS (Arhar and Gorjanc, 2007), which has been carefully sampled and contains all the most frequent genres and text types. It had also already been part-of-speech tagged and lemmatized.

For French, we used the 150 million-word corpus of newspaper articles for French Est Républicain[2] (Nehbi and Gaiffe, 2009; Seddah *et al.*, 2012), which consists of text data corresponding to two years of all the complete editions of the regional daily. Since it had not been preprocessed, we tagged and lemmatized it ourselves, using the MElt tagger and lemmatizer (Denis and Sagot, 2008).

[2] http://www.cnrtl.fr/corpus/estrepublicain/

3. Method and Experiments

The method we used for cleaning our wordnets can be divided in two steps:

1. Co-occurrence-based evaluation of the similarity between each nominal occurrence in a large (monolingual) corpus and their possible synsets as provided by the input wordnet;
2. Global assessment of all nominal (literal, synset) pairs based on these similarity measures.

Note that in the work described here, we have restricted our search for outliers to nominal synsets only, and we did not take into account multiword literals. This means that, at the time beeing, we consider as content words all tokens tagged as nouns, verbs, adjectives and adverbs, ignoring multi-word lexical units.

3.1 Basic co-occurrence-based scoring of (literal, synset) pairs

In order to achieve step 1, we first associate each synset from the input wordnet (WOLF or sloWNet) with a set of *related synsets*, i.e., a subset of all synsets (nominal or not) that are related to the base synset by relation paths of length 0, 1 or 2, based on manually designed relation patterns shown in Table 1. Second, we associate each synset pair with the list of its *related literals*, i.e. all literals that belong to any of its related synsets. We refer to the set of literals related to a synset s in the input wordnet as $R(s)$.

Path of length 0 *each synset s is related to itself*
Paths of length 1 $s \xrightarrow[\substack{hyponym\\instance\ hyponym\\mero\ portion\\mero\ part\\mero\ member}]{} s_r$
Paths of length 2 $s \xrightarrow[\substack{hyponym\\instance\ hyponym\\mero\ member\\mero\ part\\mero\ portion}]{} \xrightarrow[\substack{hyponym\\instance\ hyponym\\mero\ member\\mero\ part\\mero\ portion}]{} s_r$ $s \xrightarrow[\substack{eng\ derivative\\holo\ member\\holo\ part\\holo\ portion\\hypernym\\instance\ hypernym}]{} \xrightarrow[\substack{eng\ derivative\\holo\ member\\holo\ part\\holo\ portion\\hypernym\\instance\ hypernym}]{} s_r$ $s \xrightarrow[\substack{hypernym\\instance\ hypernym}]{} \xrightarrow[\substack{hyponym\\instance\ hyponym}]{} s_r$

Table 1: Relation paths starting from a synset s and leading to its *related synsets* s_r.

Next, given an occurrence of a nominal literal in the corpus, we look at all content words that co-occur in the same paragraph. We score each corresponding (literal, synset) pairs as follows: each literal that is related to the synset and co-occurs in the same paragraph as the occurrence increases the score by its number of occurrences divided by the number of different synsets it appears in. This gives less importance to highly polysemous literals. The computed similarity score is then normalized by dividing it by the number of content words in the paragraph.

More formally, let l be a nominal literal in paragraph p. We refer to the set of all synsets containing a word w in the input wordnet as $S(w)$, and to the number of such synsets $|S(w)|$. For example, $S(l)$ is the set of all synsets containing the nominal literal l. Let $C(p)$ be the set of (POS-tagged) content words in paragraph p, and $occ(w,p)$ the number of occurrences of the content word w in p. Finally, let $length(p)$ be the number of tokens in p. Each (literal, synset) pair of the form (l,s), with $s \in S(l)$, receives for paragraph p a local score $local\_score(l,s,p)$ defined as follows:

$$local\_score(l,s,p) = \frac{1}{length(p)} \sum_{w \in C(p) \cap R(s)} \frac{occ(w,p)}{|S(w)|}$$

The corpus-wide score $global\_score(l,s)$ for the (literal, synset) pair (l,s) is then simply the sum of the local scores of each of its occurrences:

$$global\_score(l,s) = \sum_{p} occ(l,p) \cdot local\_score(l,s,p)$$

Let us illustrate this on an example. Consider the English noun *question* ('question' in French). It appears in as many as 12 synsets in WOLF:

- eng-30-07196682-n {*question, interrogative, interrogative sentence, interrogation*}; the French *question* is correct in this synset; related literals: *examiner, interroger, phrase, question* ('examine', 'ask', 'sentence', 'question');
- eng-30-11410625-n {*outcome, consequence, upshot, effect, event, issue, result*}; the French *question* is not correct in this synset; excerpt of the related literals: *aboutir, amener, cause, chance, changement, conduire, consequence, danger, donner, découler*… ('lead to', 'lead to', 'cause', 'luck', 'change', 'lead to', 'consequence', 'danger', 'give', 'follow'…).

In our corpus, the French noun *question* occurs 26,629 times. The global score for the correct (*question*, eng-30-07196682-n) pair, based on the above-mentioned related literals, is only 182, whereas that for the incorrect (*question*, eng-30-11410625-n) pair it is as high as 1710. But at this stage, global scores do not allow us to correctly detect the erroneous (literal, synset) pair.

3.2 Extracting outlier candidates for (literal, synset) pairs

At this stage, we have for each (literal, synset) pair a global score that is the sum of the local scores of its occurrences in the corpus. We first normalize this global scores by dividing it by the sum $synset\_global\_score(s)$ of the global scores of all (literal, synset) pairs involving the same synset s. This is used to assess the *contribution* of a given literal among all literals in s. Let us call $L(s)$ the set of all literals that belong to the synset s in the input wordnet. We define $synset\_global\_score(s)$ in a straightforward way:

$$synset\_global\_score(s) = \sum_{l \in L(s)} global\_score(l,s)$$

The contribution of l to the synset s is then:

$$contribution(l,s) = \frac{global\_score(l,s)}{synset\_global\_score(s)}$$

This contribution is then normalized by the number of occurrences $occ(l)$ of the literal in the corpus, thus leading to the final score for the (literal, synset) pair (l,s):

$$score(l,s) = \frac{contribution(l,s)}{occ(l)}$$

If we go back to the example given in Section 3.1, the synset global score for eng-30-07196682-n (which means that *question* is its only literal) is 182, and is as high as 10788 for eng-30-11410625-n. Their respective contributions are thus 1 and 0.16. Our last formula then leads to a score of $37 \cdot 10^{-5}$ for (*question*, eng-30-07196682-n), whereas the score for (*question*, eng-30-11410625-n) is below $0.6 \cdot 10^{-5}$. Our final score now correctly identifies the correct vs. the incorrect (literal, synset) pairs.

4. Results and evaluation

The result of our experiment is a set of (literal, synset) pairs for each language. Each (literal, synset) pair in the list is associated with a score, all literals being necessarily attested in the corpus. We obtained 22,002 such pairs for French and 37,356 for Slovene (the difference between these figures is due at least in part to the difference in corpus sizes and genres).

Because the two corpora we have used are different in many respects (different language, different notion of what is a paragraph, different genres), we did not expect scores to be comparable between both languages. Therefore, we empirically defined two separate thresholds that define the minimum score under which a (literal, synset) pair is considered as a candidate outlier.

As mentioned above, the overall error rate in WOLF and sloWNet has been evaluated at respectively 14% and 15%, i.e., around 7,000 and 13,000 incorrect (literal, synset) pairs respectively. Therefore, we have chosen thresholds such that the number of candidate outliers has the same order of magnitude than the estimated number of erroneous (literal, synsets) pairs. This led us to thresholds of respectively $2 \cdot 10^{-5}$ and $4 \cdot 10^{-6}$ for French and Slovene, generating respectively 7,392 and 12,578 candidate outliers, i.e., approximately one third of all (literal, synset) pairs in our results.

We manually evaluated a random sample of 100 candidate outliers from each input wordnet, namely WOLF for French and sloWNet for Slovene. Among these candidates, the proportion of (literal, synset) pairs which have correctly been detected as errors is as high as 67% for French and 64% for Slovene. These figures can be compared with the estimated overall error rates in the input wordnets (12% and 15% respectively). The results are therefore very satisfying, and will be manually validated in the next months, thus leading to cleaner wordnets.

Examples of candidate outliers from WOLF (French) and sloWNet (Slovene) extracted from our manual evaluation data are shown in Table 2. Apart from the synset and the literal, we indicate the corresponding score as well as the outcome of the manual evaluation in which the 'OK' label means that the (literal, synset) pair has been correctly detected as incorrect, while the 'NO' label means that the (literal, synset) pair is indeed correct, and that its detection as a candidate outlier is erroneous.

5. Conclusions and future work

In this paper we have shown how erroneous synset candidates can successfully be eliminated from wordnets that were created by translating synsets from one language to another. The main contribution of the paper is an automated approach to clean noisy wordnets that were generated by translating synsets from a source-language wordnet into a target language via various existing bilingual resources.

The presented approach could very well be used for identifying outliers in wordnets that were constructed from monolingual resources as well, be it with clustering of words from reference corpora or from explanatory dictionaries.

In the future we plan to extend the technique to other parts of speech, and to refine it so that it would also be able to deal with subtler cross-lingual polysemy issues, which are another major cause of noise in automatically generated synsets.

Acknowledgments

The work described in this paper has been funded in part by the French-Slovene PHC PROTEUS project "Building Slovene-French linguistic resources: parallel corpus and wordnet" (22718UC), by the French national grant EDyLex (ANR-09-CORD-008) and by the Slovene national postdoctoral grant (Z6-3668).

Candidate outliers found in WOLF					Candidate outliers found in sloWNet				
literal	**synset id**	**English literal in the synset**	**score ($\times 10^3$)**	**eval**	**literal**	**synset id**	**English literals in the synset**	**score ($\times 10^3$)**	**eval**
abord	08307589	*meeting, group meeting*	0.013	OK	*aktiva*	05154517	*plus, asset*	0.002	OK
activité	14006945	*activeness, action, activity*	0.014	NO	*cilj*	05868477	*end*	0.004	OK
activité	05833022	*business*	0.011	OK	*dan*	15113229	*period, period of time, time period*	0.001	NO
adresse	00035189	*achievement, accomplishment*	0.017	OK	*dan*	15157225	*day*	0.004	NO
agence	03015254	*chest, chest of drawers, bureau, dresser*	0.015	OK	*dan*	06210791	*light*	0.003	OK
besogne	06545137	*deed of conveyance, title, deed*	0.012	OK	*dan*	06832572	*n, N*	0.004	OK
bout	08566028	*terminal, end*	0.019	NO	*datelj*	15159583	*date, day of the month*	0.000	OK
bureau	13945102	*office, power*	0.006	OK	*del*	05867413	*division, part, section*	0.003	NO
cadre	10069645	*executive director, executive*	0.017	OK	*del*	13809207	*constituent, component, component part, part, portion*	0.003	NO
cadre	10014939	*managing director, manager, director*	0.014	OK	*delež*	05256358	*part, parting*	0.004	OK

Table 2: Example of manually evaluated candidate outliers. We show the first 10 pairs in the evaluation data set for each resource, which was randomly extracted from the full sets of candidate outliers

References

Agirre, E., Ansa, O., Arregi, X., Arriola, J. M., Diaz de Ilarraza, A., Pociello, E. and Uria, L. (2002). Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. *In Proceedings of the first International WordNet Conference* in Mysore, India, 21-25 January 2002.

Arhar, Š. and Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa (The FidaPLUS corpus: a new generation of the Slovene reference corpus). *Jezik in slovstvo*, 52(2).

Bond, F., Isahara, H., Kanzaki, K. and Uchimoto, K. (2008). Boot-strapping a WordNet using Multiple Existing WordNets. *In Proceedings of LREC-2008*, Marrakech.

Fišer, D. and Sagot, B. (2008). Combining multiple resources to build reliable wordnets. In *Proceedings of TSD'08*, Brno, Czech Republic.

Lin, D. Zhao, S., Qin, L. and Zhou, M. (2003). Identifying Synonyms among Distributionally Similar Words. *In Proceedings of IJCAI'2003*, pp. 1492-1493.

Mihalcea, R., Sinha, R. and McCarthy D. (2010). SemEval-2010 Task 2: Cross-Lingual Lexical Substitution *In Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations ACL 2010*, Uppsala, Sweden.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Nehbi, K. and Gaiffe, B. (2009). TEI Est Républicain: Encodage du corpus TEI P5. Available at http://www.cnrtl.fr/corpus/estrepublicain/est-documentation.php

Pascal, D. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. *In Proceedings of PACLIC 2009*, Hong-Kong, China.

Sagot, B. and Fišer, D. (2012). Automatic extension of WOLF. In *Proceedings of the 6th Global WordNet Conference*, Matsue, Japan (to appear).

Seddah, D., Candito, M., Crabbé, B. and Anguiano, E. H. (2012). Ubiquitous Usage of a French Large Corpus: Processing the Est Republicain Corpus. *In Proceedings of LREC 2012*, Istanbul, Turkey.

Widdows, D. and Ferraro, K. (2008). Semantic vectors: a scalable open source package and online technology management application. In *Proceedings of LREC'08*, Marrakech, Morocco.